

Universal and Distinct Properties of Communication Dynamics: How to Generate Realistic Inter-event Times

Pedro O.S. Vaz de Melo, Universidade Federal de Minas Gerais

Christos Faloutsos, Carnegie Mellon University

Renato Assunção, Universidade Federal de Minas Gerais

Rodrigo Alves, Universidade Federal de Minas Gerais

Antonio A.F. Loureiro, Universidade Federal de Minas Gerais

With the advancement of information systems, means of communications are becoming cheaper, faster and more available. Today, millions of people carrying smart-phones or tablets are able to communicate at practically any time and anywhere they want. Among others, they can access their e-mails, comment on weblogs, watch and post comments on videos, make phone calls or text messages almost ubiquitously. Given this scenario, in this paper we tackle a fundamental aspect of this new era of communication: how the time intervals between communication events behave for different technologies and means of communications? Are there universal patterns for the inter-event time distribution (IED)? In which ways inter-event times behave differently among particular technologies? To answer these questions, we analyze eight different datasets from real and modern communication data and we found four well defined patterns that are seen in all the eight datasets. Moreover, we propose the use of the Self-Feeding Process (SFP) to generate inter-event times between communications. The SFP is extremely *parsimonious* point process that requires at most two parameters and is able to generate inter-event times with all the universal properties we observed in the data. We show the potential application of SFP by proposing a framework to generate a synthetic dataset containing realistic communication events of any one of the analyzed means of communications (e.g. phone calls, e-mails, comments on blogs) and an algorithm to detect anomalies.

Categories and Subject Descriptors: H.2.8 [Information Systems]: database management—*Database Applications, Data mining*; G.3 [Mathematics of Computing]: Probability and Statistics—*Statistical computing*

General Terms: Theory

Additional Key Words and Phrases: communication dynamics, inter-event times, generative model

1. INTRODUCTION

A popular saying that came with the advancement of information systems is that the distance among people is decreasing over the years. It is well known that the main reason for that is the fact that means of communications are becoming cheaper, faster and more available. Today, millions of people carrying smart-phones or tablets are able to communicate at practically any time and anywhere they want. Among others, they can access their e-mails, comment on weblogs, watch and post comments on videos, make phone calls or text messages almost ubiquitously. It is fascinating that the growing accessibility, reach and speed of these means of communications are making them more and more homogeneous and similar. For instance, consider a smart-phone user with a permanent Internet connection. What is the fastest way to reach this person? By a phone call, by a SMS message, by e-mail or by an instant messaging subscription service (e.g. WhatsApp)? Nowadays, maybe all of these are equally or similarly effective.

Given this scenario, in this paper we tackle a fundamental aspect of this new era of communication: how the time intervals between communication events behave for different technologies and means of communications? Are there universal patterns for the inter-event time distribution (IED)? In which ways inter-event times behave differently among particular technologies? To answer these questions, we analyze eight different datasets from real and modern communication data, that can be divided into two groups. The first group contains five datasets extracted from Web applications in which several users comment on a given topic. The datasets are extracted from five popular websites: Youtube, MetaFilter, MetaTalk, Ask MetaFilter and Digg. The second group contains three datasets in which individuals perform and receive communication events. In this group we have a Short Message Service (SMS), a mobile phone-call and a public e-mail dataset. These datasets comprise a set of different types of interactions that are common and routine in most human lives.

As the first contribution of this paper, we found four well defined patterns that are seen in all the eight datasets. First, we show that the marginal distribution of the time intervals between communications follows an odds ratio power law. Second, we show that the slope of this power law is approximately 1 for the majority of the data analyzed. Third, unlike previous studies, we analyze the temporal correlations between inter-event times, illustrating the “i.i.d. fallacy” that has been routinely ignored until recently [Karsai et al. 2012]. We show that, unlike the PP that generates independent and identically distributed (i.i.d.) inter-event times, individual sequences of communications tend to show a high dependence between consecutive inter-arrival times. Finally, we show that the collection of individual IEDs of all systems is very well modeled by a Bivariate Gaussian Distribution. Moreover, besides these four universal properties, we also identified features that differentiate one system from the other and that naturally come from the idiosyncrasies of each system.

As the second contribution of this paper, we propose the use of the Self-Feeding Process (SFP) to generate inter-event times between communications of an individual or in blog posts or videos. The SFP is extremely *parsimonious* point process that requires at most two parameters. We show that it is able to generate inter-event times with all the universal properties we observed in the data and also reconciles existing and contrasting theories in human communication dynamics [Barabási 2005; Malmgren et al. 2008]. Moreover, we show how the SFP can be easily modified to also encompass the particularities seen in each of the analyzed systems.

Finally, as the third contribution of this paper, we show two possible applications of the findings described in this paper. First, through the use of the SFP, we propose a framework to generate a synthetic dataset containing realistic communication events of any one of the analyzed means of communications (e.g. phone calls, e-mails, comments on blogs). This framework considers all the universal properties and the particularities of each system. Second, we show how to detect anomalies in the systems we investigated through the use of this framework. Among the regular individuals, we were able to identify, for instance, a SMS automated service, blog posts that were deleted by the moderators because of their content and a polemic Youtube video populated by flaming¹ discussions.

The rest of the paper is organized as follows. Section 2 provides a brief survey of the related work that analyzed inter-event times between communications. Section 3 describes the eight datasets used in this work. Section 4 shows the IED of individuals from these datasets and that the Odds Ratio function of their IEDs is well modeled by a power law. Section 5 shows that the typical behavior of inter-event sequences shows a positive correlation between consecutive inter-event times. Section 6 describes the SFP model, which provides an intuitive and simple explanation for the observed data. Section 7 shows that the SFP model also unifies existing theories on communication dynamics. Section 8 describes a model to represent the collective behavior of users in the analyzed systems. Section 9 shows a method to spot anomalies. Finally, we show the conclusions and future research directions in Section 10.

2. RELATED WORK

The study of the time interval in which events occur in human activity is not new in the literature. The most primitive model is the classic Poisson process [Haight 1967]. Although the most recent approaches have among themselves significant differences, they all agree that the timing of individuals systematically deviates from this classical approach. The Poisson process predicts that the time interval Δ_t between two consecutive events by the same individual follows an exponential distribution with expected value β and rate $\lambda = 1/\beta$, where

$$\Delta_t = -\beta \times \ln(U(0, 1)), \quad (1)$$

where $U(0, 1)$ is a uniformly random distributed number between $[0, 1]$. While in a Poisson process consecutive events follow each other at a relatively regular time, real data shows that humans have very long periods of inactivity and also bursts of intense activity [Barabási 2005].

¹hostile and insulting interaction between Internet users

Moreover, recent analysis on the time interval between communication activities shows apparent conflicting ideas among them. First, Barabási [Barabási 2005] proposed that bursts and heavy-tails in human activities are a consequence of a decision-based queuing process, when tasks are executed according to some perceived priority. In this way, most of the tasks are rapidly executed and some of them may take a very long time. The queuing models proposed in [Barabási 2005] generates power law [Faloutsos et al. 1999] distributions $p(X = x) \approx x^{-\alpha}$ with slopes $\alpha \approx 1$ or $\alpha \approx 1.5$. In the literature, there are examples that are approximated by the universality class model in e-mail records [Eckmann et al. 2004; Vazquez et al. 2006], web surfing [Dezsö et al. 2006; Vazquez et al. 2006], library visitation, letters correspondence and stock broker's activities [Vazquez et al. 2006], arrival times of requests to print in a student laboratory [Harder and Paczuski 2006] and in short-messages [Wei et al. 2009], most of them reporting slopes from 1 to 1.5 and, in the case of [Wei et al. 2009], also slopes higher than 1.5. Although a power law visually fits well the tail of the IED, it usually can not explain the whole distribution [Malmgren et al. 2008].

Second, other work in literature propose that the IED is well explained by variations of the PP, such as the Interrupted Poisson [Kuczura 1973] (IPP), Non-Homogeneous Poisson Process [Malmgren et al. 2008], Kleinberg's burst model [Kleinberg 2002] and others. For instance, Malmgren et al. [Malmgren et al. 2008] proposed a non-homogeneous Poisson process to explain the inter-event times distribution. The model is based on the circadian and weekly cycles and coupled to the cascading activity and has a varying rate $\lambda(t)$ that depends on time t in a periodic manner. This process generates active intervals according to $\lambda(t)$. Each active interval initiates a homogeneous Poisson process with a determined rate λ_a . In order to generate the active intervals, the model needs (i) the average number of active intervals per week, and (ii) the probabilities of starting an active interval at a particular time of day and (iii) week. Malmgren et al. estimated these parameters empirically and they showed that the model accurately fits the real data. However, this model explains the data at the cost of requiring several parameters and careful data analysis, being impractical for synthetic data generators, for instance. Later, the authors adapted this model to a more parsimonious version [Malmgren et al. 2009a], but it still has 9 parameters.

3. DATA DESCRIPTION

In this work we analyze eight datasets that can be divided into two groups. The first group contains five datasets extracted from Web applications in which several users comment on a given topic. The datasets are extracted from five popular websites: Youtube, MetaFilter, MetaTalk, Ask MetaFilter and Digg. The second group contains three datasets in which individuals perform and receive communication events. In this group we have a Short Message Service (SMS), a mobile phone-call and a public e-mail dataset. For simplicity, we use the term "individual" to refer both to topics of the first group and users of the second group.

In the first group, we analyze a public online news dataset, containing a set of stories and comments over each story. More specifically, the data is from the popular social media site Digg and has 1,485 stories and over 7 million comments [De Choudhury et al. 2009]. The Digg dataset is public for research interests and can be downloaded at <http://www.infochimps.com/datasets/diggcom-data-set>. We also analyze three publicly available datasets from the *Metafilter Infodump Project*², extracted from three discussion forums: MetaFilter³ (Mefi), MetaTalk⁴ (Meta) and Ask MetaFilter⁵ (Askme). After disregarding topics which received less than 30 comments, the Mefi dataset has 8,384 topics and 1,471,153 comments, the Meta dataset has 2,484 topics and 503,644 comments and the Askme dataset has 498 topics and 65,950 comments.

²downloaded on September 22nd from <http://stuff.metafilter.com/infodump/>

³<http://www.metafilter.com/>

⁴<http://metatalk.metafilter.com/>

⁵<http://ask.metafilter.com/>

Our final dataset from the first group was collected from the Youtube website using the Google's Youtube API⁶. We collected all the comments posted on the videos classified as *trending* by the API⁷ from 22/Aug/2012 to 25/Sep/2012. We collected a total of 1,221,390 comments on 989 videos, but we use in our dataset only those videos with more than 30 comments and which the comments span for more than one week, a total of 610 videos and 1,008,511 comments. The full dataset can be downloaded at www.dcc.ufmg.br/~olmo/youtube.zip.

In the second group, the mobile phone calls dataset contains more than 3.1 million customers of a large mobile operator of a large city, with more than 263.6 million phone call records registered during *one month*. From this same operator, we also have a SMS dataset of 300,000 users spanning six months of data, for a total of 8,784,101 records. These datasets from the mobile operator is under Non-Disclosure Agreement (NDA) and belong to the iLab Research at the Heinz College at CMU, but was already used in several papers [Vaz de Melo et al. 2010; Vaz de Melo et al. 2011; Akoglu et al. 2012]. We also analyze the public Enron e-mail dataset, consisting of 200,399 messages belonging to 158 users with an average of 757 messages per user [Klimt and Yang 2004]. The data is public and can be downloaded at <http://www.cs.cmu.edu/~enron/>.

4. MARGINAL DISTRIBUTION

In this work, we are first interested on the inter-event time distribution IED of the random variable Δ_k representing the time Δ_k between the k -th and the $(k-1)$ -th communication events on a given topic (first group) or of an user (second group). For simplicity, we use the term “individual” to refer both to topics (videos, blog posts, news) of the first group and users of the second group.

4.1. Odds Ratio Using the Cumulative Distribution Function

In Figure 1, we show the distribution of the time intervals Δ_k between communication events for a typical active user of the SMS dataset, with 44785 SMS messages sent or received. The histogram is showed in Figure 1-a and, as we observe, this user had a significantly high number of events separated by small periods of time and also long periods of inactivity. Moreover, both the power law fitting, which in the best fit has an exponent of -2 , and the exponential fitting, which is generated by a PP, deviates from the real data. The method we use to fit the power law is based on the Maximum likelihood estimation (MLE) described in [Clauset et al. 2009].

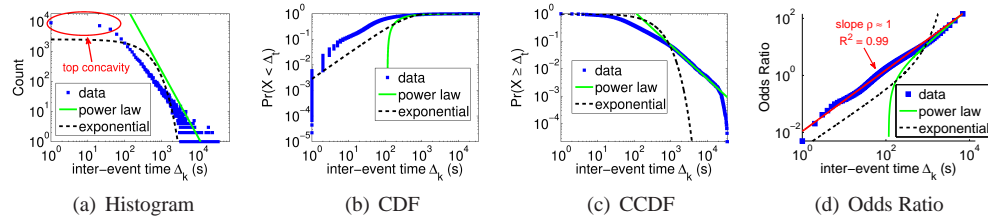


Fig. 1. The inter-event times distribution of the most talkative user of our four datasets, with 44785 SMS messages sent and received. We observe that both the power law fitting (PL fitting) with exponent ≈ 2 and the exponential fitting, generated by a PP, deviate from the real data. We also observe that the OR is very well fitted by a straight line with slope ≈ 1 .

In empirical data that spans for several orders of magnitude, which is the case of the IEDs, it is very difficult to identify statistical patterns in the histograms, since the distribution is considerably noisy at its tail [Barabási 2005; Malmgren et al. 2008]. A possible option is to move away from the histogram and analyze the cumulative distributions, i.e., cumulative density function (CDF) and complementary cumulative density function (CCDF), which veil the data sparsity. However, by using the CDF, as we observe in Figure 1-b, we lose information in the tail of the distribution and,

⁶<https://developers.google.com/youtube/>

⁷https://gdata.youtube.com/feeds/api/standardfeeds/on_the_web

on the other hand, by using the CCDF, as we observe in Figure 1-c, we lose information in the head of the distribution.

In order to escape from these drawbacks, we propose the use of the Odds Ratio (OR) function combined with the CDF as it allows for a clean visualization of the distribution behavior either in the head or in the tail. This $OR(k)$ function is commonly used in the survival analysis [Bennett 1983; Mahmood 2000] and measures the ratio between the number of individuals who have not survived by time t and the ones that survived. Its formula is given by:

$$Odds\ Ratio(t) = OR(t) = \frac{CDF(t)}{1 - CDF(t)}. \quad (2)$$

In this paper, for a set of n inter-event times $\{\Delta_1, \Delta_2, \dots, \Delta_n\}$, we calculate the odds ratio for each percentile P_1, P_2, \dots, P_{100} of the data. This avoids that minor deviations in the data harms the goodness of fit test we perform, which we explain in Section 4.2.

Thus, in Figure 1-d, we plot the OR for the selected user. From the OR plot, we can clearly see the cumulative behavior in the head and in the tail of the distribution. Also, observe again that both the exponential and the power law significantly deviate from the real data. Moreover, we can also observe that the OR of the inter-event times seems to entirely follow a linear behavior in logarithmic scales. That is, $\log(OR(t))$ is a linear function of $\log(t)$ with slope $\rho \approx 1$ and we say that we have a OR power law behavior. If the approximation is turned into an equality, this implies that the inter-event times follow a log-logistic distribution (see Appendix B.1).

In Figure 2, we plot the OR of a typical individual of each dataset. As in Figure 2-d, the OR plots show a clear and almost perfect linear relationship between $\log(OR(t))$ and $\log(t)$ in all the examples considered. This implies that the IED follows a log-logistic distribution, as we explain in Appendix B.1. Also, we can observe that the OR of the inter-event times seems to follow entirely the same linear behavior in logarithmic scales, having, then, an OR power law behavior. This implies that the marginal distribution of the IEDs is approximately equal to a log-logistic distribution [Fisk 1961], since it also shows a OR power law behavior. For a larger sample, please see the Appendix E.

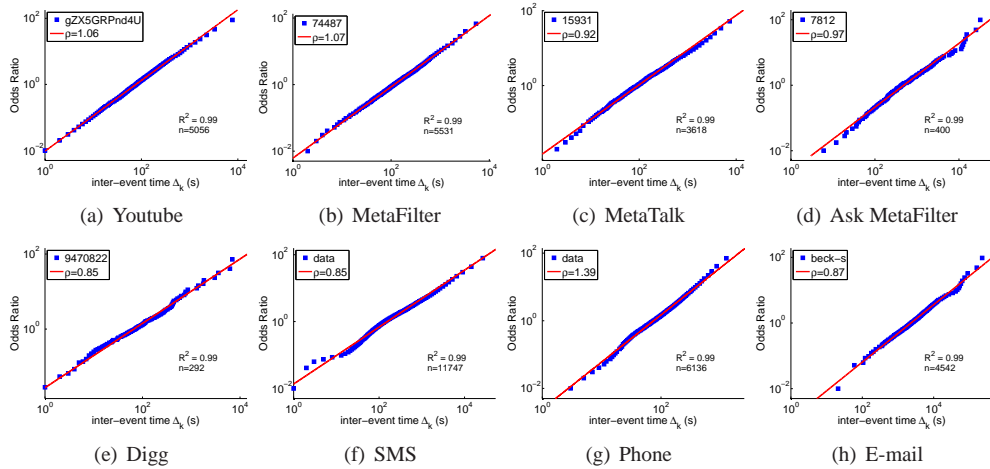


Fig. 2. The Odds Ratio plot for one typical active individual of each dataset. Observe that an odds ratio power law, represented by a straight line with slope ρ in a log-log scale, is an appropriate fit for all individuals.

4.2. Goodness of Fit

In this section, we check whether the OR of the IEDs of all individuals of our datasets can be explained by a power law. We perform a linear regression using least squares fitting on the OR of

the IEDs of all individuals. Since we consider every percentile and the OR is based on the CDF, a cumulative distribution, the linear regression can be used to measure the goodness of fit. We performed a Kolmogorov-Smirnov goodness of fit test, but because of digitalization errors and other deviations in the data, this test is only approximated..

Figure 3 shows the histogram of the determination coefficient R^2 of the performed linear regressions. The determination coefficient R^2 is a statistical measure of how well the regression line approximates to the real data points. An $R^2 = 1.0$ indicates that the regression line perfectly fits the data. We observe that for the vast majority of individuals of our eight datasets, the R^2 is very close to 1.0. More specifically, in the first group, the R^2 averages 0.99 for the phone dataset, 0.96 for the SMS dataset and 0.97 for the e-mail dataset. For the second group, the R^2 averages 0.97 for the Youtube, Askme and Digg datasets and 0.98 for the Mefi and Meta datasets. This allows us to state the following universal pattern:

UNIVERSAL PATTERN 1. *The Odds Ratio of the inter-event time distribution of communication events is well fitted by a power law.*

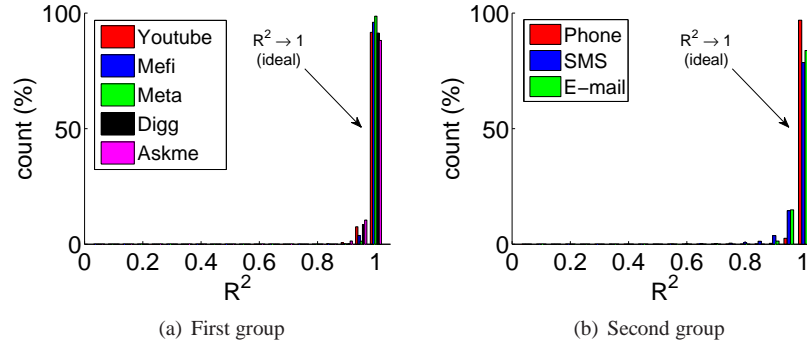


Fig. 3. The goodness of fit of our proposed model. We show the histograms of the R^2 s measured for every user in the eight datasets. These histograms consider bins of size 0.05. Thus, observe that the R^2 value for the great majority of individuals is located in the last bin, from 0.95 to 1.

4.3. Odds Ratio of Well Known Distributions

We have seen that the IED of the majority of the individuals of our datasets is well modeled by an odds ratio power law. In Figure 4, observe that this odds ratio power law behavior is also seen in log-logistically distributed data and cannot be seen in other well known distributions. This is the first indication that the marginal distribution of time intervals between communications of individuals is well modeled by a log-logistic distribution.

5. TEMPORAL CORRELATION

Although most previous analysis focus solely on the marginal IED, a subtle point is the *correlation* between successive inter-event times (Δ_{k-1} and Δ_k). What we illustrate here is that the independence between Δ_k and Δ_{k-1} does **not** hold for the eight datasets we analyzed in this work.

In Figure 5, we plot, for the same typical users of Figure 2, all the pairs of consecutive inter-event times (Δ_{k-1}, Δ_k). We also show the regression of the data points using the LOWESS smoother [Cleveland 1979]. While the PP, as for any other renewal process, the regression is a flat line with slope 0, for the eight typical users Δ_k tends to grow with Δ_{k-1} . This means that if I called you five years ago, my next phone call will be in about five years later. In short, there is a strong, positive dependency between the current inter-event time (Δ_k) and the previous one (Δ_{k-1}), clearly contradicting the independence assumption.

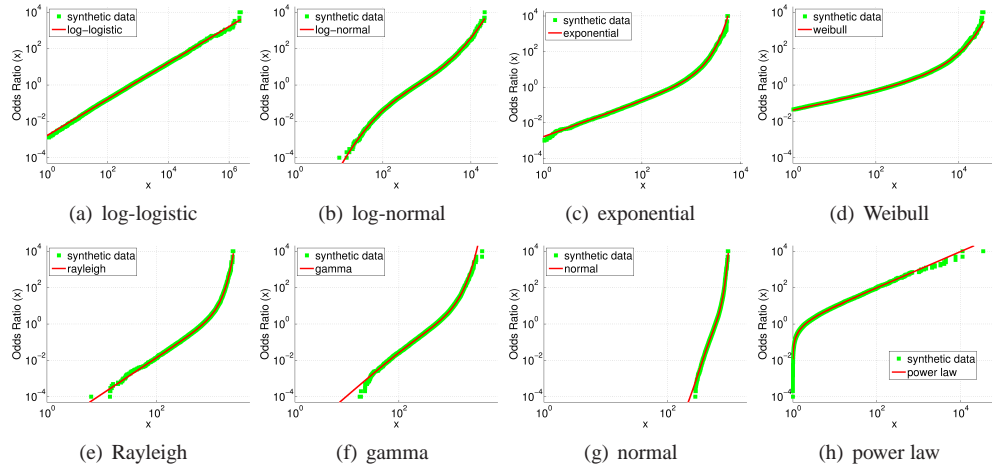


Fig. 4. The Odds Ratio function for eight well known distributions.

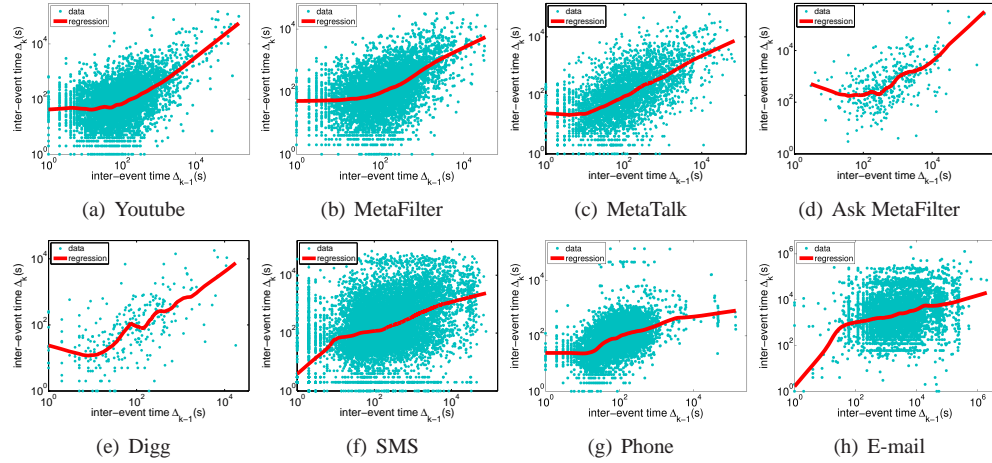


Fig. 5. I.i.d. fallacy: dependence between Δ_k and Δ_{k-1} . Each point represents a pair of consecutive inter-event times (Δ_{k-1}, Δ_k) registered for a typical active individual of each dataset. The red line is a regression of the data points using the LOWESS smoother.

We formally investigate if two consecutive inter-event times are correlated analyzing the autocorrelation [Box et al. 1994] of all the time series involving the inter-event times Δ_k of the individuals of our datasets. Autocorrelation refers to the correlation of a time series with its own past and future values. A positive autocorrelation, which is suggested by Figure 5, might be considered a specific form of “persistence”, i.e., a tendency for a system to remain in the same state from one observation to the next.

We test if all the Δ_k time series of every individual of our datasets are random or autocorrelated. For this, we define the hypothesis test H_0 that a series $S = \{\Delta_0, \Delta_1, \dots, \Delta_n\}$ of inter-event times is random. If S is random, then its empirical autocorrelation coefficient $AC_l \approx 0$ for all lags $l > 0$, where a lag l is used to compare, in this case, values of Δ_k and Δ_{k-l} . More formally, if AC_l is within the 95% confidence interval for S to be random, then we accept H_0 that S is random. As we show in Figure 6, we reject the null hypothesis H_0 that the inter-event times of the individual of Figure 1 is random, since all AC_l , $1 < l \leq 10$ are outside the confidence interval.

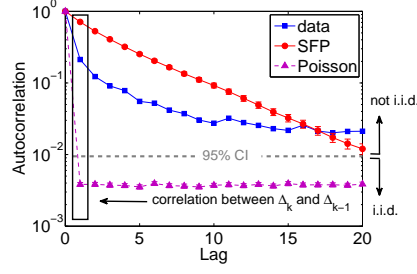


Fig. 6. The sample autocorrelation for the same individual of Figure 1 and for synthetic data generated by our proposed SFP model (see Section 6) and a PP with the same number of communication events and median.

Since we are interested only in the case where the lag $l = 1$, we propose an alternative hypothesis test H_1 that the first-order autocorrelation coefficient AC_1 is greater than 0. If AC_1 is greater than the confidence interval for randomness, then we accept H_1 that the series is not random, i.e., there is a dependence between Δ_k and Δ_{k-1} . In Figure 7, we show the proportion $P(H_1)$ of individuals in our data to which H_1 is true grouped by their number of events n . As we observe, as the number of communication events n grows and becomes significant, $P(H_1)$ increases rapidly. This strongly suggests that, on the contrary of what happens with the i.i.d. inter-event times distribution generated by the Poisson Process or simply sampling from a log-logistic distribution (LLG-iid), in real data there is a dependence between Δ_k and Δ_{k-1} . This also agrees with a recent work [Owczarczuk 2011], which reports that daily series of number of calls made by a customer exhibits strong autocorrelation. Thus, in summary we can state that

$$E(\Delta_k) \neq E(\Delta_k | \Delta_{k-1}) \quad (3)$$

or

$$E(\Delta_k | \Delta_{k-1}) = f(\Delta_{k-1}) \quad (4)$$

where f is a function that describes the dependency between Δ_k and Δ_{k-1} . Moreover, we can state the following universal pattern:

UNIVERSAL PATTERN 2. *There is a significant positive correlation between two consecutive inter-event times.*

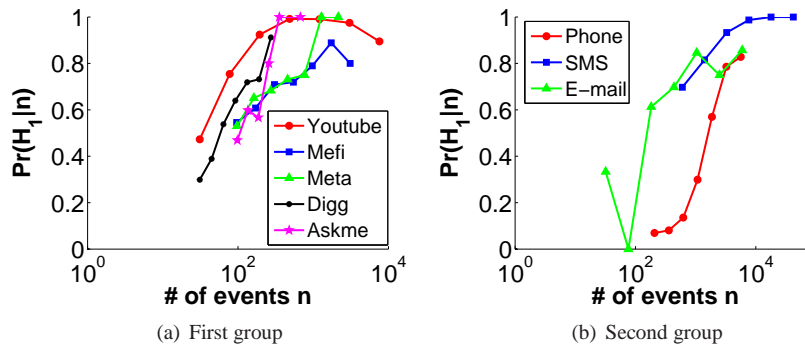


Fig. 7. The proportion $P(H_1)$ of individuals in our data to which H_1 is true grouped by their number of events. Note that as the number of events grows, the proportion of individuals that have autocorrelated series increases rapidly for the eight datasets.

6. THE SELF-FEEDING PROCESS

Given all the above evidence (OR power law; i.i.d. fallacy) and all the previous evidence (power law tails by Barabási; short-term regular behavior as the PP), the question is whether we can design a generator which will match all these properties? Our requirements for the ideal generator are the following:

- R1: Realism – marginals.* The model should generate OR power law marginal IED;
- R2: Realism – locally-Poisson.* The model should behave as a Poisson Process within a short window of time;
- R3: Avoid the i.i.d. fallacy.* Two consecutive inter-event times should be correlated;
- R4: Parsimony.* It should need only few parameters, and ideally, just one or two.

6.1. Candidate Parameters

Since the IED of the majority of individuals is well modeled by an odds ratio power law, which implies a log-logistic distribution, we can characterize their behavior by the two parameters, the slope ρ and the median μ , from the linear relationship in the OR plot. Observe in Figure 8 the PDF of the slopes ρ_i measured for every individual i of our eight datasets. Except the SMS dataset, the typical ρ_i for the majority of individuals is approximately 1. This surprising result allows us to state the following universal pattern:

UNIVERSAL PATTERN 3. *The typical slope of the odds ratio power law that best fits the inter-event time distribution of an individual is 1.*

Moreover, observe in Figure 9 the PDF of the medians μ_i measured for every individual i of our eight datasets. Observe that, while the typical μ_i is around 1 hour for the first group, for the second group it varies from 3 to 8 minutes. Thus, in Section 6.2 we propose a simplified one-parameter model that generates IEDs with slopes $\rho = 1$ and varied medians. Then, in Section 6.3, we propose a generalized two-parameter model that generates IEDs with varied slopes and medians.

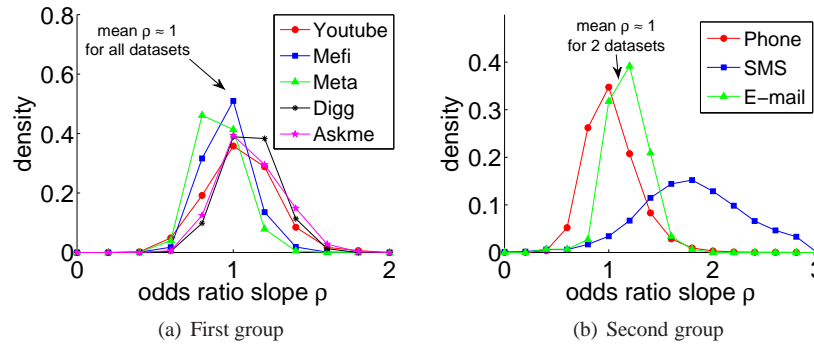


Fig. 8. The PDF of the slopes ρ_i measured for every user u_i of our eight datasets. Except the SMS dataset, the typical ρ_i for the majority of individuals is approximately 1.

6.2. The Simplified SFP Model

At a high level, our proposal is that the next inter-arrival time will be an exponential random variable, with rate that *depends on the previous* inter-arrival time. It is subtle, but in this way our generator behaves like Poisson in the short term, gives power-law tails in the long term, generates OR power law marginals and is extremely parsimonious: just one parameter, the median μ of the IED. We call this model the *Self-Feeding Process* (SFP).

We propose the generator as follows

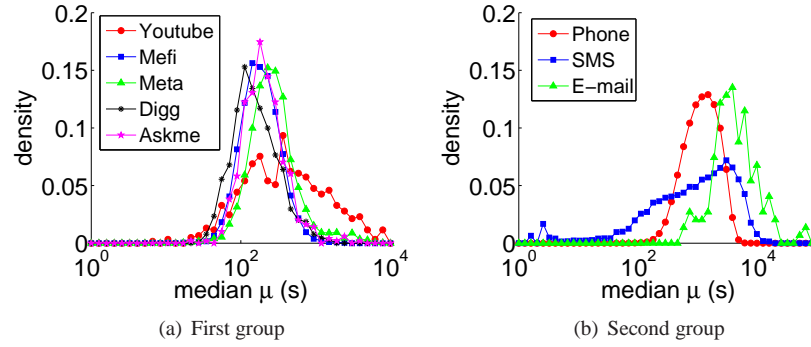


Fig. 9. The PDF of the medians μ_i measured for every user of our eight datasets. Observe that, while the typical μ_i is around 3 and 8 minutes for the first group, and around 1 hour for the second group.

MODEL 1. *Self-Feeding Process SFP* (μ).
 $//\mu$ is the desired median of the marginal PDF

$$\begin{aligned} \Delta_1 &\leftarrow \mu \\ \Delta_k &\leftarrow \text{Exponential}(\text{mean } \beta = \Delta_{k-1} + \mu/e) \end{aligned}$$

where μ is the only parameter of the model, being the desired median of the IED. The part μ/e must be greater than 0 to avoid Δ_k to converge to 0 and has to be divided by the Euler's number e to make the median of the generated IED around the target median μ (more details in the Appendix A). This type of model is not new in the literature [Wold and universitet. Statistiska institutionen 1948; Cox 1955] but they have not been extensively studied, perhaps due to the lack of empirical data fitting the implied distribution.

In Figures 10-a and 10-b we compare, respectively, the histogram and the OR of the inter-event times generated by the SFP model, all values rounded up, with the inter-event times of the individual of Figure 1. Notice that the distributions are very similar and both are well fitted by a log-logistic distribution, which looks like a hyperbola, thus addressing both the power-law tail, as well as the “top-concavity” that real data exhibits.

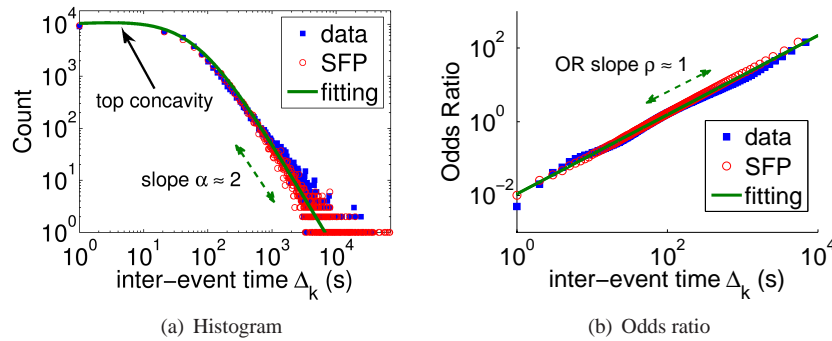


Fig. 10. Comparison of the marginal distribution of the inter-event times generated by the SFP model with the inter-event times of the most active user of Figure 2. Observe that both the histogram (a) and the OR (b) are almost identical.

The SFP model naturally generates an odds ratio power law for IED with slope $\rho = 1$, which is the slope that characterizes the majority of the users of our datasets (see Figure 8). To the best

of our knowledge, this is the first work that studies the IED of human communications using such a varied, modern and large collection of data. Despite the fact that the means of communications are intrinsically different, having their own idiosyncrasies, we have observed that the IED of most individuals of these systems have the same characteristics, i.e., they follow an odds ratio power law behavior. Moreover, when the OR slope $\rho = 1$, the power law exponent of the PDF is $\alpha = -2$ (see the Appendix B.6 for details). This is the same IED slope α reported in [Hidalgo 2006; Vazquez et al. 2006] as a result of fluctuations in the execution rate and in particular periodic changes. It has been argued that seasonality can only robustly give rise to heavy-tailed IEDs with exponent $\alpha = 2$.

6.3. The Generalized SFP Model

In Figure 8 we showed the slopes ρ of the OR fitting for the IEDs of all individuals of our datasets. It is fascinating that the typical ρ_i for the individuals of seven of our datasets is approximately 1, the same slope generated by the simplified SFP model. Several individuals though, mainly from the SMS dataset, have a much higher value of ρ , close to $\rho \approx 2$. To accommodate that and all the variance seen in the data, we introduce our Generalized SFP model, which needs just one parameter more, ρ . Thus, we have:

MODEL 2. *Generalized Self-Feeding Process SFP*(μ, ρ).

$\begin{aligned} \delta_1 &\leftarrow \mu \\ \delta_t &\leftarrow \text{Exponential}(\text{mean: } \beta = \delta_{t-1} + \mu^\rho / e) \\ \Delta_k &\leftarrow \delta_t^{1/\rho}. \end{aligned}$

Note the auxiliary variable δ_t , which stores the inter-event times without the influence of ρ . For more details about the SFP parameters, please see Appendix A.

7. THE UNIFYING POWER OF THE SFP

In this section, we emphasize the unifying power of the SFP. Several works [Karagiannis et al. 2004; Malmgren et al. 2008; Malmgren et al. 2009b; Malmgren et al. 2009a; Kuczura 1973; Kleinberg 2002] claim that in the short term, real data behave as regular as a PP. Our model also captures that, since successive inter-event times are exponentially distributed, with similar (but not identical) rates. Thus, one of the major contributions of this work is the unification of the two seemingly-conflicting viewpoints we mentioned earlier. The proposed SFP model unifies both theories by generating Poisson-like traffic in the short term, with smoothly varying rate, like the second viewpoint, and also generates a power-law tail distribution (see the Appendix B.6), even matching the top-concavity that power laws can not match, like the first modern approach of Barabási [Barabási 2005].

In Figure 11, we explicitly show the SFP's unifying power. We compare synthetic data generated by the SFP model using the same odds ratio slope ρ , median μ and number of events of the user of Figure 1-a with the real data from this user. Notice the bursts of activity and also the long periods of inactivity, in the first two columns of Figure 11. Also notice that both synthetic and real traffic significantly deviate from Poisson (sloping lines in Figures 11-b and 11-f) but are similar between themselves. However, in the short term, both real and synthetic data behave like Poisson, being practically on top of the black dashed lines of Figures 11-d and 11-h.

8. COLLECTIVE BEHAVIOR

Since we know that the great majority of users' IED can be modeled by the SFP model, we can figure out how each individual i is distributed in its population according to their parameters ρ_i and μ_i of the SFP model. If the meta-distribution of the parameters ρ_i and μ_i is well defined, then we can model the collective behavior of the individuals, which may serve for various applications, such as synthetic generators, anomaly detection, among others. From now on, we will call the meta-distribution of the parameters ρ_i and μ_i the *MetaDist* distribution.

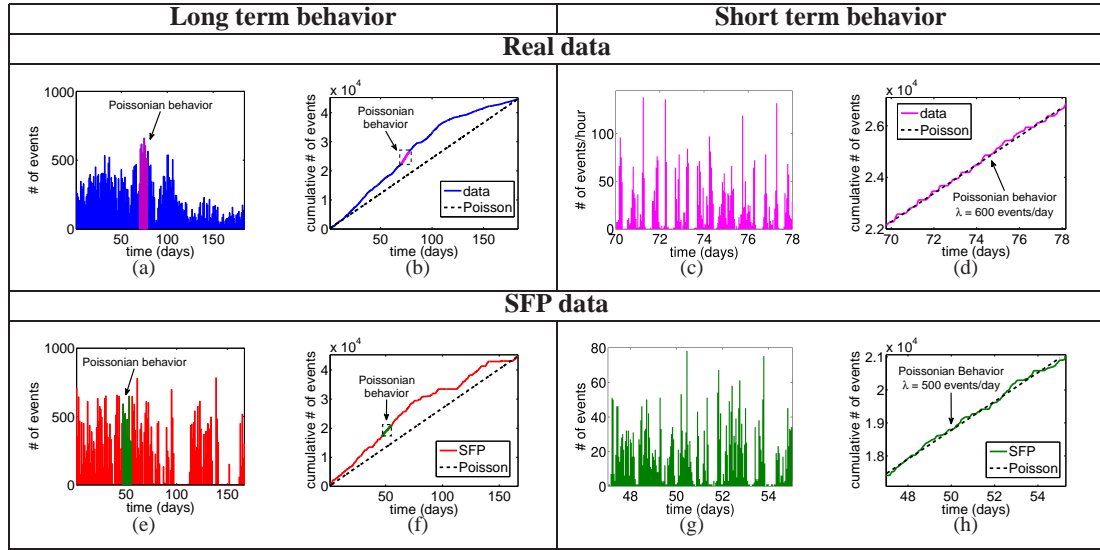


Fig. 11. Unification Power of SFP: non-Poisson/bursty in the long term, but Poisson in the short term. Real data: Traffic of the user of Figure 2-1, showing event-count per unit time (a and c) and respective cumulative event-count (b and d). SFP data: synthetic traffic generated by the SFP model (with matching μ , ρ and event-count). Observe that (1) both time series are visually similar; (2) both are bursty in the long run (spikes; inactivity) (3) both are Poisson-like in the short term (last two columns)

In Figure 12-a, we show the scatter plot of the fitted parameters ρ_i and $\log(\mu_i)$ of every MetaTalk individual i . It is difficult to visualize the patterns due to the large number of overlapping data points but, however, we can spot outliers. Moreover, by plotting the ρ_i and $\log(\mu_i)$ parameters using isocontours, as shown in Figure 12-b, we automatically smooth the visualization by disconsidering low populated regions. While darker color mean a higher concentration of pairs ρ_i and $\log(\mu_i)$, white color mean that there is small probability of observing a user with these values of ρ_i and $\log(\mu_i)$. We use $\log(\mu_i)$ instead of μ_i because, as we see in Figure 9, the logarithm of the medians can be approximated by a normal distribution for all datasets.

Surprisingly, we observe that the isocontours of Figure 12-b are very similar to the ones of a bivariate Gaussian. In order to verify this, we extracted from the *MetaDist* distribution the means P and B of the parameters ρ_i and $\log(\mu_i)$, respectively, and also the covariance matrix Σ . We use these values to generate the isocontours of a bivariate Gaussian distribution and we plotted it in Figure 12-c. We observe that the isocontours of the generated bivariate Gaussian distribution are very similar to the ones from the *MetaDist* distribution. Thus a bivariate Gaussian distribution fits the real data of fitted ρ_i s and $\log(\mu_i)$ s and hence, it is a good model to represent the population of individuals whose IED can be modeled by the SFP.

In order to verify if this pattern replicates in the other datasets, we show in Figure 13 the comparison between the real data and the synthetic data generated from the 'meta-fitting' of the parameters ρ_i and $\log(\mu_i)$. Observe that the bivariate Gaussian distribution can also model the collective behavior for all the other seven datasets, allowing us to state the following universal pattern:

UNIVERSAL PATTERN 4. *The joint distribution of the parameters ρ_i and $\log(\mu_i)$ associated with individual i of a particular communication system follows a bivariate Gaussian distribution.*

This result is very useful, since it allow us to easily generate a synthetic dataset for a particular system. In order to do that, we simply have to perform the following steps:

- (1) Select from Table I the system which you would like to generate the synthetic dataset;

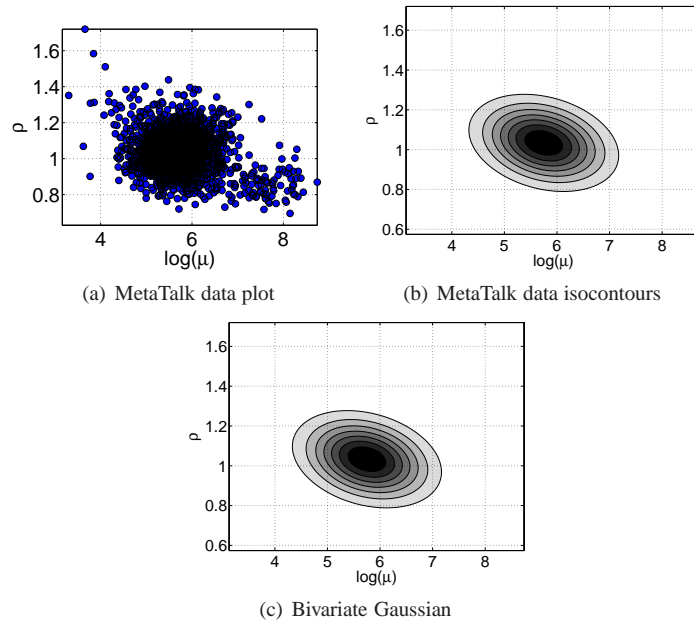


Fig. 12. Plots of the parameters ρ_i and $\log(\mu_i)$ of every MetaTalk individual i . In (a) we can not see any particular pattern, but we can spot outliers. By plotting the isocontours (b), we can observe how well a bivariate Gaussian (c) fits the real distribution of the pairs $(\rho_i, \log(\mu_i))$ ('meta-fitting').

- (2) Create a bivariate Gaussian sampler using the correspondent parameters, which are shown in Table I;
- (3) Sample the n individuals from the bivariate Gaussian sampler;
- (4) Select the duration window T of the dataset, e.g. $T = 1$ month;
- (5) For each individual i , generate N_i inter-event times $\Delta_0, \Delta_1, \dots, \Delta_{N_i}$, where N_i is the highest natural number that satisfies $\sum_k^{N_i} \Delta_k < T$.

Table I. Parameters of the bivariate Gaussian distributions.

System	AskMe	Digg	Enron	Meta	Mefi	Phone	SMS	Youtube
$E(\rho_i)$	0.927	0.930	0.830	1.004	1.033	1.388	0.920	1.023
$E(\log(\mu_i))$	5.625	5.126	8.251	5.455	5.748	5.714	5.672	5.274
$\text{Var}(\rho_i)$	0.016	0.013	0.006	0.012	0.014	0.007	0.006	0.015
$\text{Var}(\log(\mu_i))$	0.470	0.291	0.417	0.317	0.487	0.041	0.301	1.163
$\text{Cov}(\rho_i, \log(\mu_i))$	-0.028	-0.010	-0.018	0.000263	-0.021	-0.002	-0.025	-0.080

9. ANOMALIES

In the previous section, we showed that the majority of individuals of our eight datasets is well modeled by the SFP. Moreover, we showed that the collective behavior of the individual IEDs of these datasets is well modeled by a bivariate Gaussian distribution. A natural application of these findings would be for anomaly detection. An individual that does not have a IED that can be explained by the SFP is a potential individual to be observed, since it has a distinct communication behavior from the majority of the other users. Moreover, an individual i whose ρ_i and $\log(\mu_i)$ values are significantly different from the typical individual is also a likely target to investigate.

Thus, we define anomalies those individuals that fall into one of the following three criteria:

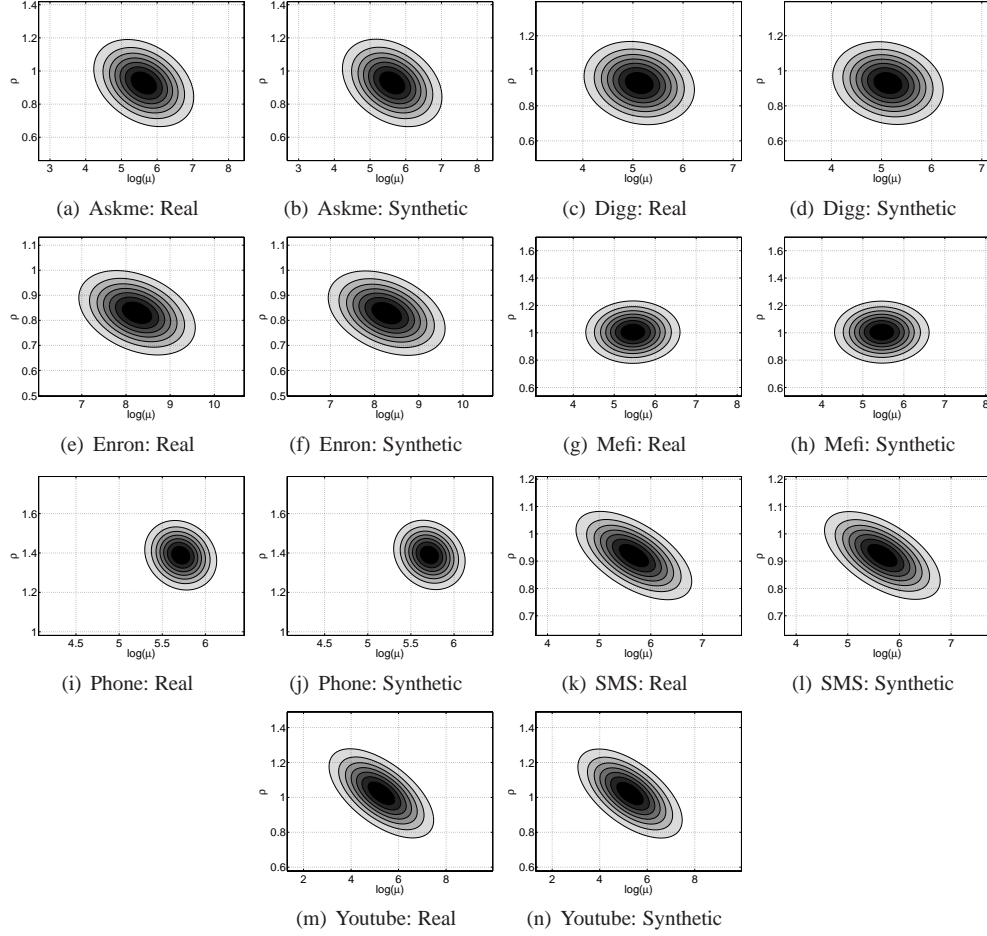


Fig. 13. Comparison between the real and synthetic datasets generated from bivariate Gaussians.

- **A1**: the IED is well modeled by SFP but its ρ_i or $\log(\mu_i)$ values are significantly distant from the bivariate Gaussian that describes the population;
- **A2**: the IED is not well modeled by the SFP but its ρ_i and $\log(\mu_i)$ values are inside the bivariate Gaussian that describes the population;
- **A3**: the IED is not well modeled by the SFP and its ρ_i or $\log(\mu_i)$ values are significantly distant from the bivariate Gaussian that describes the population.

The distance between an individual and the expected typical behavior modeled by the bivariate Gaussian distribution is calculated by the Mahalanobis distance D^2 that is commonly used to measure the distance between an individual sample point y and its expected value m [Johnson and Wichern 2007]. If y follows a bivariate Gaussian distribution with covariance matrix Σ then D is given by

$$D^2 = (y - m)^t \Sigma^{-1} (y - m)$$

which follows a chi-squared distribution with 2 degrees of freedom. Individuals with D^2 greater than 25 occurs at a rate of 4 per million and hence they are considered outliers. These anomalies are represented in Figure 14 by the colors green (A1), red (A2) and blue (A3), respectively. White individuals have or are similar to the typical behavior.

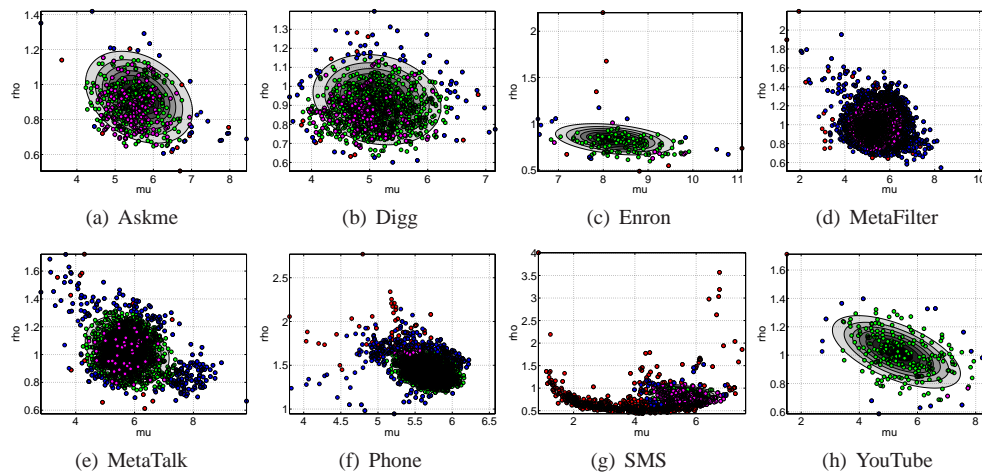


Fig. 14. Types of anomalies detected

By checking the top anomalies according to our criteria, we found interesting examples. In AskMe dataset, for example, a point out of the bivariate Gaussian (type A3) was detected by the AskMe staff as a topic containing inappropriate content for this kind of service. When we accessed the post (see Figure 15) we found the following message: “This post was deleted for the following reason: Historical outliers notwithstanding, this is not what askme is for.”. As another example, we spotted a type A1 anomaly in the MetaFilter dataset that was also detected manually by the administrator and deleted from the community. The post was changed (see Figure 15) and the following message was posted: “This post was deleted for the following reason: if you are going to be subtle, you need to be clearer.”. Considering the anomalies found in the MetaTalk dataset, we detected a type A2 anomaly that is a post from the developers talking about improvements and asking comments and suggestions to the community was spotted. This type of post is not the goal of MetaTalk. Finally, a YouTube video with the subject “Bill Nye: Creationism Is Not Appropriate For Children” was identified as a type A3 anomaly by our framework. First, while the typical number of views of the videos posted by the owner of this video is around dozens of thousands, this particular video has close to five million views. Moreover, by analyzing the content of the comments posted on this video, we could identify constant flaming, i.e., insulting interactions among the comments.



Fig. 15. Examples of deleted posts detected as anomalies for the Askme (top) and MetaFilter (bottom) datasets.

Concerning the SMS dataset, we found several interesting anomalies of the type A3. These anomalous behavior is derived from the fact that it is common that users subscribe to automated ap-

plications which periodically send messages to them about a given topic, e.g. news, movies, sports etc. These messages are counted in the IED of the user as a regular SMS or e-mail message, but they do not represent a social interaction. Thus, when the percentage of these messages is high, the IED shape is modeled by two random variables, the one which represents the social interactions and the one which represents the incoming messages from automated services. If the amount of messages of the second type is significant, then the distribution deviates significantly from the one generated by the SFP, as we observe in Figure 16.

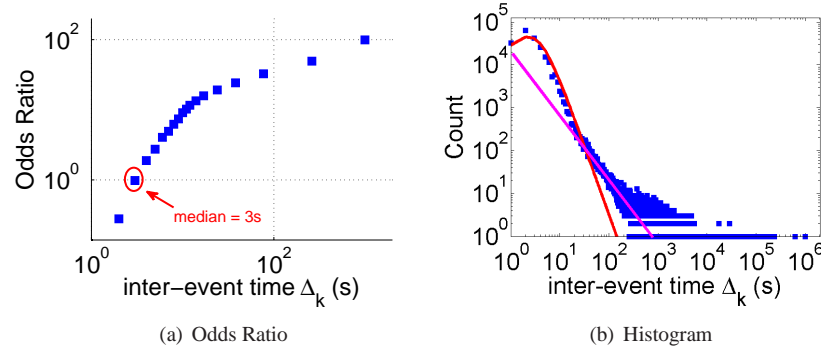


Fig. 16. The IED of the most active user of our 8 datasets, with 229590 communication events. Observe how the behavior differs from the SFP or any other well known distribution, probably due to the significant number of automatic messages sent, that represents 97% of the total number of communication events.

10. CONCLUSIONS

In this paper, we showed that eight different systems share four common properties in their communication dynamics. These universal properties are:

- (1) the marginal distribution of the time intervals between communications follows an odds ratio power law;
- (2) the slope of this power law is typically 1;
- (3) individual sequences of communications tend to show a high dependence between consecutive inter-arrival times;
- (4) the collection of individual IEDs is very well modeled by a Bivariate Gaussian Distribution.

Moreover, we proposed the SFP model, which reconciles previous approaches for human communication dynamics and also is able to generate communication events that match all the four universal properties listed above. Finally, we showed that from the knowledge presented in this paper is possible to generate realistic synthetic datasets of communications and to spot anomalies.

APPENDIX

A. PARAMETERS

Before reaching the generalized SFP model described in the paper, we had a simpler version of it, relying on a different parametrization scheme:

MODEL 3. *Self-Feeding Process SFP* (C, a).

$$\begin{aligned} \delta_1 &\leftarrow C \\ \delta_t &\leftarrow \text{Poisson Process}(\beta = \delta_{t-1} + C) \\ \Delta_k &\leftarrow \delta_t^a, \end{aligned} \tag{5}$$

where C is the location parameter and a is the shape parameter that defines the odds ratio slope ρ . An easy and direct way to define the relationships between this model's parameters and the distribution properties μ and ρ is through simulations.

Thus, the first point we consider is the median μ of the inter-event times generated by the SFP model when $a = 1$. When $OR(x) = 1$, x is the median μ of the distribution. Thus, in Figure 17-a, we plot the OR for different values of C . We observe that changing the value of C changes μ and, consequently, the location of the distribution, but maintains its slope. We also see that μ is close but different than the value of C .

In order to investigate the relationship between C and μ , we run simulations of the model for all integer values of C between $[1, 10000]$. As we observe in Figure 17-b, the median μ of the inter-event times distribution (IED) varies linearly with C according to a slope of ≈ 2.72 , that can be approximated by Euler's number e , in a way that $\mu \propto e \times C$. This allows us to generate inter-event times with a determined μ when the slope $\rho = 1$. We ignore the constant factor 3.8 because its 95% confidence interval is $(-8.596, 16.3)$, which contains zero.

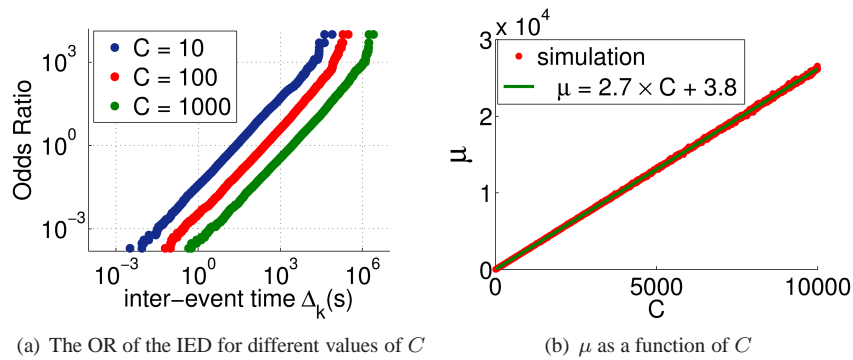


Fig. 17. Changing the value of C changes the location of the distribution. The median of the distribution μ varies linearly with C , $\mu = a \times C + b$, with $a = 2.719$ and $b = 3.8$. The 95% confidence interval for a is $(2.715, 2.723)$ and for b is $(-8.60, 16.3)$. Since the confidence interval for b contains 0, b is not significant.

Now we know how to generate inter-event times with different medians μ using the parameter $C = \mu/e$ of SFP. The next step is to verify how the SFP model can generate IEDs with a desired slope $\rho \neq 1$. Considering that up to this point the SFP model generates a set of inter-event times I_1 with a slope 1, the idea is to use an exponent a to transform I_1 into I_ρ , which is an IED with a different slope ρ . When we elevate each $\Delta_k \in I_1$ to the power of $a \neq 1$, the resulting slope ρ becomes different from 1, as we see in Figure 18-a. In the same way we did for C , we run simulations of the model for 1000 different values of $a \in [0.1, 2]$. As we observe in Figure 18-b, there is an inverse relationship between a and ρ , i.e., $\rho = a^{-1}$. Moreover, since the median of the distribution is also elevated to the power of a , we have to elevate the parameter μ to the power of $\rho = a^{-1}$ to preserve the median.

Regarding the need for the constant μ/e in SFP, we propose the following lemma:

LEMMA A.1. *The constant $C = \mu/e > 0$ of Model 2 is needed to assure that the inter-event times generated by the SFP model will not converge to zero.*

PROOF. If we remove the constant C from Model 2, $\Delta_k = (\Delta_{k-1}) \times (-\ln(U(0, 1)))$, or Δ_k will be equal to Δ_{k-1} multiplied by a random number X extracted from the exponential distribution with parameter $\beta = \lambda = 1$. If $(X = \frac{1}{k} \mid k > 1)$, then Δ_k will be equal to Δ_{k-1} divided by k . The probability of X to be $\frac{1}{k}$ is $P(X = \frac{1}{k}) = e^{-\frac{1}{k}} = \frac{1}{k^k e}$. On the other hand, the probability of multiplying Δ_k by k and, therefore, return Δ_{k+1} to Δ_{k-1} value is $P(X = k) = e^{-k} = \frac{1}{e^k}$. Given

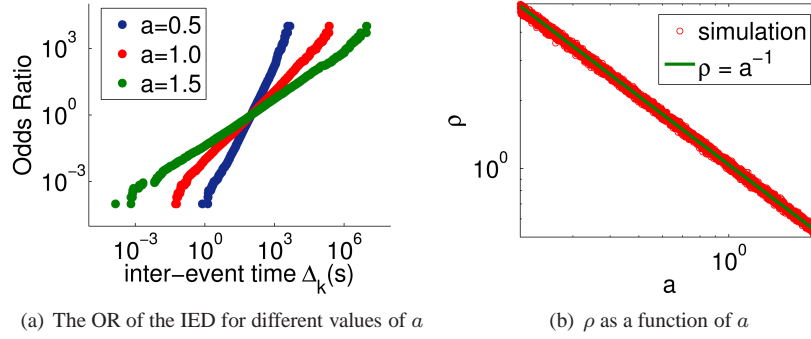


Fig. 18. Changing the value of a changes the slope ρ of the distribution in a way that $\rho = a^{-1}$.

these probabilities, observe that $P(X = \frac{1}{k}) = \frac{1}{k^{\frac{1}{\rho}}e} > P(X = k) = \frac{1}{k^{\frac{1}{\rho}}e}, \forall k > 1$. From this, we conclude that the expected value of Δ_k when $t \rightarrow \infty$ is 0. With C in the equation, even when $\Delta_{k-1} = 0$, $\Delta_k = -C \times \ln(U(0, 1))$, that is a classic Poisson process with $\beta = C$, and, obviously, does not converge to 0. \square

B. THE SFP STATIONARY DISTRIBUTION

In this section, we analyze the properties of the stationary distribution generated by the SFP. All the results shown in this section are coherent with the following conjecture:

CONJECTURE B.1. *The SFP model generates a log-logistic distribution with $\rho = 1$,*

where $\rho = 1/\sigma$ and σ is the shape parameter of the log-logistic distribution.

As we show in this section, we have several and significant evidences that the SFP generates a log-logistic distribution, but at this moment we do not have a formal analytical proof that this is true.

B.1. Log-logistic Distribution

The log-logistic distribution was first proposed by Fisk [Fisk 1961] to model income distribution, after observing that the OR plot of real data in log-log scales follows a power law $OR(x) = cx^\rho$. In summary, a random variable is log-logistically distributed if the logarithm of the random variable is logistically distributed. The logistic distribution is very similar to the normal distribution, but it has heavier tails. In the literature, there are examples of the use of the log-logistic distribution in survival analysis [Bennett 1983; Mahmood 2000], distribution of wealth [Fisk 1961], flood frequency analysis [M.I. Ahmad and Werritty 1988], software reliability [Gokhale and Trivedi 1998] and phone calls duration [Vaz de Melo et al. 2010]. A commonly used log-logistic parametrization is [Lawless and Lawless 1982]:

$$\begin{aligned} PDF_{LLG}(x) &= \frac{e^z}{\sigma x(1+e^z)^2}, \\ CDF_{LLG}(x) &= \frac{1}{1+e^{-z}}, \\ z &= (\ln(x) - \ln(\mu))/\sigma, \end{aligned} \quad (6)$$

where $\sigma = 1/\rho$, the slope of our SFP model, and μ is the same. Moreover, when $\sigma = 1$, it is the same distribution as the Generalized Pareto distribution [Lorenz 1905] with shape parameter $\kappa = 1$, scale parameter μ and threshold parameter $\theta = 0$.

B.2. Analytical Result

Wold processes [Wold and universitet. Statistiska institutionen 1948; Cox 1955] are stochastic processes where the inter-events intervals have a dependence following a Markovian property. That

is, the probability distribution law of the t -th inter-event time Δ_t depends only on the previous inter-event time Δ_{t-1} . Our SFP model falls within this Wold processes class as we assume that, conditionally on the entire previous inter-event times, the distribution of $\Delta_k = \delta_t^{1/\rho}$ is an exponential distribution with expected value given by $\delta_{t-1} + \mu^\rho/e$. Wold processes are not well understood due to the mathematical difficulties in deriving their probabilistic properties.

Consider the existence of a stationary distribution for the generalized SFP model. A stationary PDF $f(x)$ of the Markov chain δ_t must satisfy

$$\begin{aligned} f(x) &= \int_0^\infty f(y \rightarrow x) f(y) dy \\ &= \int_0^\infty \frac{1}{y + \mu^\rho/e} \exp(-x/(y + \mu^\rho/e)) f(y) dy \end{aligned}$$

This integral equation has no obvious analytical solution but in the next sections we show via simulations of the point process that $f(x)$ is very well approximated by a log-logistic density. This mathematical difficulty is common in the previous attempts to model data with Wold processes. Even if a consistent density $f(x)$ and a transition kernel $f(y \rightarrow x)$ are given, properties are, in general, difficult to obtain [Cox and Isham 1980].

Let the Markovian distribution of δ_t conditional on $\delta_{t-1} = x$ be given by an exponential distribution with mean $\alpha x + c$ where $0 < \alpha \leq 1$ and $c > 0$ are constants. We have

$$\mu_t = \mathbb{E}(\delta_t) = \mathbb{E}\{\mathbb{E}(\delta_t|\delta_{t-1})\} = \mathbb{E}\{\alpha\delta_{t-1} + c\} = \alpha\mu_{t-1} + c$$

Applying recursively, we find

$$\mu_t = \alpha^t \mu_0 + c \sum_{k=0}^{t-1} \alpha^k$$

If $\alpha = 1$, $\mu_t = \mu_0 + tc$. Assuming that this process is stationary implies that $\mu_t = \mu_0$ is constant and the only solution is to take $\mu_t = \mu_0 = \infty$. Hence the process has infinite mean, as is the case of the log-logistic distribution with shape parameter equal or smaller than 1.

If $\alpha < 1$, then

$$\mu_t = \alpha^t \mu_0 + c \frac{1 - \alpha^t}{1 - \alpha} \rightarrow \frac{c}{1 - \alpha}$$

Obviously, $\mu_t = \mu = c/(1 - \alpha)$ is a solution to the recursive equation $\mu_t = \alpha\mu_{t-1} + c$.

When we have a finite expectation for δ_t we can calculate the variance $\mathbb{V}(\delta_t) = \sigma_t^2$:

$$\begin{aligned} \sigma_t^2 &= \mathbb{E}\{\mathbb{V}(\delta_t|\delta_{t-1})\} + \mathbb{V}\{\mathbb{E}(\delta_t|\delta_{t-1})\} \\ &= \mathbb{E}\{(\alpha\delta_{t-1} + c)^2\} + \mathbb{V}\{\alpha\delta_{t-1} + c\} \\ &= \alpha^2 (\sigma_{t-1}^2 + \mu_{t-1}^2) + 2c\alpha\mu_{t-1} + c^2 + \alpha^2 \sigma_{t-1}^2 \\ &= 2\alpha^2 \sigma_{t-1}^2 + (\alpha\mu_{t-1} + c)^2 \end{aligned}$$

Assuming that the process is stationary, we have $\mu_t = c/(1 - \alpha)$ and $\sigma_t^2 = \sigma^2$ constant, which implies into

$$0 < \sigma^2 = 2\alpha^2 \sigma^2 + (c/(1 - \alpha))^2$$

If $\alpha < 1/\sqrt{2}$, this has a solution as

$$\begin{aligned}\sigma^2 &= \left(\frac{c}{1-\alpha} \right)^2 \frac{1}{(1-\sqrt{2}\alpha)(1+\sqrt{2}\alpha)} \\ &= \mu^2 \frac{1}{(1-\sqrt{2}\alpha)(1+\sqrt{2}\alpha)}\end{aligned}$$

Therefore, if $\alpha = 1$, the process has infinite mean. If $\alpha < 1$, the expected value is finite and equal to $\mu = c/(1-\alpha)$. Concerning the variance, it exists only if $\alpha < 1/\sqrt{2} \approx 0.70$ and, in this case, we have $\sigma^2 = \mu^2/(1-2\alpha^2)$.

B.3. Fitting Synthetic Data

In Figure 19-a, we plot the histogram of 100,000 time intervals Δ_k generated by the SFP model with $\mu = e$. Moreover, in Figure 19-b, we plot the OR for the same time intervals. While a classic PP generates an exponential distribution, we observe that the generated data by the SFP perfectly fits a distribution with an Odds Ratio function that is a power law with slope $\rho = 1$. This is also coherent with Conjecture B.1.

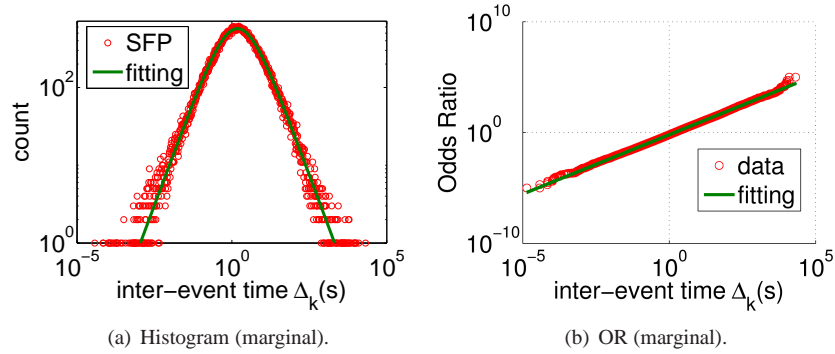


Fig. 19. Inter-event times Δ_k generated by the SFP. The generated Δ_k s are perfectly fitted by a log-logistic distribution with the slope $\rho = 1$.

B.4. The SFP Markov Chain

The SFP can be naturally considered as a Markov Chain (MC), since it is a sequence of random variables $\Delta_1, \Delta_2, \Delta_3, \dots$ with the Markov property, namely that, given the present inter-event time, or state, the future and past inter-event times, or states, are independent. Thus, here we model the SFP as a time-homogeneous Markov chain with a finite state space to give another evidence that the SFP has a stationary distribution and that is very likely that this distribution is the log-logistic.

Originally, the SFP can be considered as a continuous-time MC, but for simplicity, we build a discrete-time Markov chain in a way that each state $i = \{1, 2, 3, \dots\}$ is associated with an inter-event time $\Delta_i = \{\Delta_1, \Delta_2, \Delta_3, \dots\}$ with values within the interval $(i-1, i]$. For instance, considering the granularity in seconds, if the current inter-event time is 3.8 seconds, then the MC is in the state 4. Also for simplicity, we build a finite-state MC with a maximum number of states n , i.e., the states go from 1 to n . The MC will be in state n every time the current inter-event time is within the interval (n, ∞) .

Thus, considering a n -state MC build from the SFP model, the transitions probabilities $p_{i,j}$ of going from state i to j are given in the following way:

$$p_{i,j} = \begin{cases} CDF_{exp}(x=j, \beta=i+C) - CDF_{exp}(x=j-1, \beta=i+C) & \text{if } j < n \\ 1 - CDF_{exp}(x=j, \beta=i+C) & \text{if } j = n, \end{cases}$$

where $CDF_{exp}(x, \beta)$ is the cumulative distribution function of the exponential distribution on x with mean β and $C = \mu/e$, given in the SFP (Equation 1). Observe in Figure 20 that the probability density function of the log-logistic is virtually identical to the one of the stationary distribution of the SFP Markov Chain. This is another strong indication that the SFP generates log-logistically distributed data.

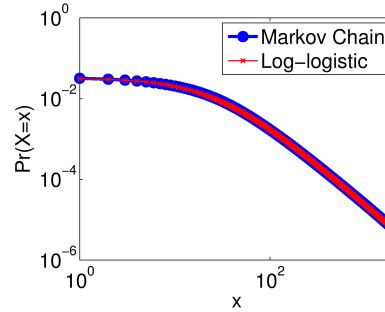


Fig. 20. The probability density function of the log-logistic distribution and the stationary distribution of the SFP MC.

It is important to point out that in [Chierichetti et al. 2012] the authors showed that behavior of the Web users is not Markovian, i.e., a user's next action does not depend only on her/his current state. Our assumption differs from this one because we assume that users have Markovian behavior in communications, while [Chierichetti et al. 2012] studied whether users have Markovian behavior while navigating on the Web.

B.5. Solving by Computation

If the SFP model generates a log-logistic distribution with slope $\rho = 1$, then we can write the PDF of the inter-event times generated by the SFP model as

$$f(x) = \int_0^\infty PDF_{LLG}(y; \rho = 1, \mu) \times PDF_{EXP}(x; \beta = y + C) dy, \quad (7)$$

or

$$f(x) = \int_0^\infty \left(\frac{1}{\mu} \times \frac{1}{\left(1 + \frac{y}{\mu}\right)^2} \right) \times \frac{e^{-\frac{x}{y+C}}}{y+C} dy. \quad (8)$$

In this way, we verify if Equation 8 is correct by numerically evaluating the integral using the adaptive Gauss-Kronrod quadrature method for every $x \in]0 : 10^4]$ and comparing the result with the $PDF_{LLG}(x)$. As we can see in Figure 21, the proposed PDF and the PDF_{LLG} match perfectly for different values of μ .

B.6. SFP has Power Law Tail

The universality class model proposed by Barabási [Barabási 2005] states that the IED has a power law tail. The proposed SFP model agrees with this model in a way that:

LEMMA B.2. *If Conjecture B.1 is correct, then the SFP model generates an IED that converges to a power law when $x \rightarrow \infty$, i.e., $\lim_{x \rightarrow \infty} \frac{PDF_{LLG}(x)}{x^{-\alpha}} = k$, where k is a constant greater than 0.*

PROOF. Considering the Probability Density Function of the log-logistic distribution showed in Equation 6, if we set the location parameter $\mu = 1$ for simplicity, $e^z = x^{1/\sigma}$. Then, $PDF_{LLG}(x)$

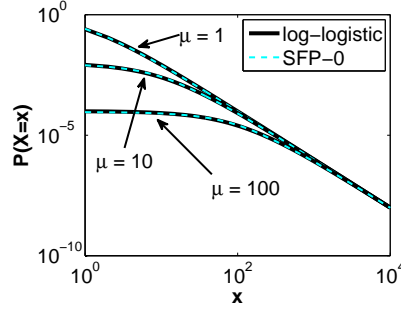


Fig. 21. The PDF of the log-logistic distribution VS. the PDF of Equation 8, that is derived from the SFP. Observe that the PDFs are identical for different values of μ . We numerically evaluated the integral using the adaptive Gauss-Kronrod quadrature method.

can be simplified to

$$PDF_{LLG}(x) = \frac{x^{\frac{1}{\sigma}-1}}{\sigma(1+x^{\frac{1}{\sigma}})^2}.$$

When $x \rightarrow \infty$, the addition of 1 in the denominator can be disregarded, resulting in the following simplification:

$$PDF_{LLG}(x) = \frac{x^{-(1+1/\sigma)}}{\sigma}, x \rightarrow \infty.$$

Thus, when $x \rightarrow \infty$, the IED generated by the SFP model is a power law with slope

$$\alpha = -(1 + 1/\sigma) = -(1 + \rho). \quad (9)$$

Observe again Figure 19-a and note the power law tail. \square

C. GENERATING REALISTIC INTER-EVENT TIMES

Time intervals between communications are very hard to model. We showed in this work that even today there is no consensus about the more appropriate model to represent them. One of the main reasons for that is that real data is significantly noisier, what may harms deeply the statistical analysis. In this section, we show some of these noisy behaviors seen in real data and, more important, we show how to reproduce them in order to generate more realistic data.

First, in Section C.1, we describe minor deviations seen in data that we prefer not to approach via changes in the SFP model. Then, in the following sections, we make slight alterations in the SFP in order to explain the most significant deviations seen in data. In Section C.2, we show how to generate inter-event times with more realistic correlation between consecutive inter-event times. Then, in Section C.3, we show how to mimic the inter-event times between SMS messages sent to multiple recipients. Finally, in Section C.4, we add an overhead parameter to the SFP model in order to explain the apparent lower bound for inter-event times seen in the phone dataset.

C.1. Minor Deviations

In most of the IEDs, there are small deviations from the OR power law for Δ_k values close to 10 hours. This is explained by the regular sleep intervals between two communication events. Since everyone has to go to sleep and it is not usual to call during the sleeping hours, it is common to have a time interval of approximately 10 hours at every 24 hours, more than what is expected by the fitting. For more details about sleep intervals, refer to [Vaz de Melo et al. 2011].

Additionally, it is also common to see a high amount of communication events at round times, such as 1 minute or 1 hour, because the recorded time of the event is rounded by the server, e.g. 8:54 is rounded to 9:00. We overcome such deviations by computing the odds ratio for the percentiles

of the distribution. Since the typical individual has thousands of communication events, then these deviations are not shown in the odds ratio plot.

Finally, it is important to consider the deviations that appear specifically in the SMS dataset, that is the dataset that presented the worst goodness of fit result. Probably the main reason for that is the fact that a significant amount of messages arrive at their destinations with a considerable delay, such as human delay and other noisy non-regular delays caused by the mobile network infrastructure or personal issues, e.g., a customer left his mobile phone unattended and the battery died, delaying all the incoming SMS messages for when the mobile phone is recharged again. Imagine, for instance, that Smith had sent a message to John at time t_1 and, due to a transmission delay d_1 , the message arrived only at t_2 . In his turn, Smith saw the message at t_2 and immediately replied, but again, due to a transmission delay d_2 , the message arrived to John only at t_3 . Thus, for John, the inter-event time between sending the message and receiving the reply is $\Delta = t_3 - t_1 = (t_2 + d_2) - (t_1 + d_1)$, with *two* transmission delays embedded in the registered inter-event time.

C.2. Lower Temporal Correlation

The SFP model is build upon a direct dependence between consecutive inter-event times. Because of that, the correlation between consecutive inter-event times is significantly higher than real data. While the average Pearson's correlation coefficient for real data is approximately 0.4, for synthetic data generated by the SFP model is approximately 0.7. In order to generate more realistic data, we suggest a slight modification in the SFP process. Instead of generating the next inter-event time (Δ_k) based on the immediate previous one (Δ_{k-1}), we propose that it should be generated from a ϵ -th previous one ($\Delta_{k-\epsilon}$). This can be done by extracting ϵ from an exponential distribution with mean $\beta = 1$ and making its ceiling, so the lower bound for ϵ is 1. In summary, the SFP model is changed as follows:

MODEL 4. *Self-Feeding Process* SFP* (μ).
 μ is the desired median of the marginal PDF

$$\begin{aligned} \Delta_1 &\leftarrow \mu \\ \epsilon &\leftarrow \lceil \text{Exponential}(\text{mean } \beta = 1) \rceil \\ \Delta_k &\leftarrow \text{Exponential}(\text{mean } \beta = \Delta_{k-\epsilon} + \mu/e) \end{aligned}$$

Observe in Figure 22 that the synthetic data generate by the SFP* has a lower correlation (0.43) between consecutive inter-event times than the original one (0.70). Despite of that, the odds ratio generated by the SFP* is still a power law with slope $\rho \approx 1$.

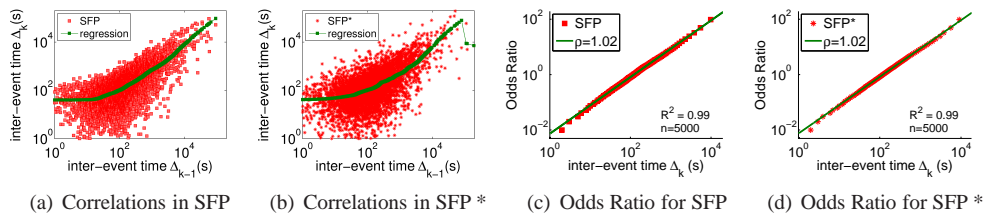


Fig. 22. Comparison between the synthetic data generated by the SFP and the SFP *. Observe that the synthetic data generate by the SFP * has a lower correlation (0.43) between consecutive inter-event times than the SFP (0.70). Despite of that, the odds ratio generated by the SFP * is still a power law with slope $\rho \approx 1$.

C.3. Multiple Recipients and Sleep Intervals

When sending SMS messages it is common to copy the message to multiple recipients. Because of that, several IEDs of the SMS dataset show a deviation from the OR power law in the first seconds of the distribution. In order to mimic this behavior, we propose two small changes in the SFP. First, we generate the number of recipients of a communication event from an exponential distribution (e.g.: mean $\beta = 1$). Then, we send the communication event to every recipient with a delay also extracted from an exponential distribution (e.g.: mean $\beta = 1$). As we observe in Figure 23, these small changes are able to represent the inter-event times sent to multiple recipients.

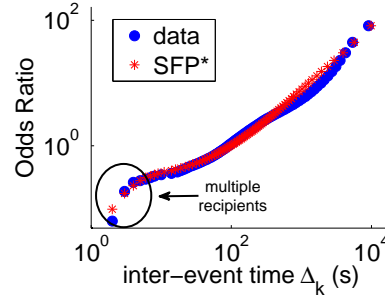


Fig. 23. The comparison of the between the IED of a typical SMS user with the one synthetically generated by the SFP * model described in this section.

C.4. Phone Dialing Speed Limit

For the majority of phone users, the number of Δ_k values close to 10 seconds is underestimated by the odds ratio power law fitting. This happens because the IED of phone data is usually lower bounded by the setup time Δ_k^0 of making a phone call, that involves dialing the numbers, waiting for the signal, and waiting for the other part to answer the call. We can mimic this behavior by simply adding a overhead constant θ to every inter-event time generated by the SFP model, representing the time it takes for a individual to dial and wait for the reply. We change the SFP model in the following way:

MODEL 5. *Self-Feeding Process* SFP* (μ, θ).

// μ is the desired median of the marginal PDF // θ is the usual time it takes for a individual to dial and get the reply

$$\begin{aligned} \Delta_1 &\leftarrow \mu + \theta \\ \Delta_k &\leftarrow \text{Exponential}(\text{mean } \beta = \Delta_{k-\epsilon} + \mu/e) + \theta \end{aligned}$$

Observe in Figure 24 that this simple modification can accurately mimic the phone dialing speed limit seen in real data.

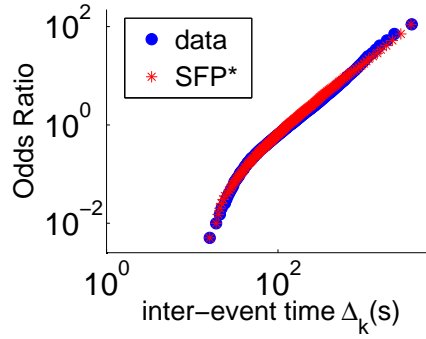


Fig. 24. SFP * vs. real data

D. SFP CODE

Below we show the *Python* code for the SFP generator.

```
def SFP(n, mu, rho=1):
    #first inter-event time
    deltat = mu
    #list of inter-event times
    Deltat = []
    for i in range(1, n):
        #Poisson Process which Beta=deltat+mu/e
        deltat = -(deltat+(mu**rho)/math.e)
        deltat = deltat * math.log(random.random())
        Deltat.append(deltat**(1/rho))
    return Deltat
```

E. DATA SAMPLE

In this section we show the IED of 8 typical talkative individuals of each dataset. In each figure we show the identification of the individual (when possible), the slope ρ , the determination coefficient R^2 and the number of communication events n .

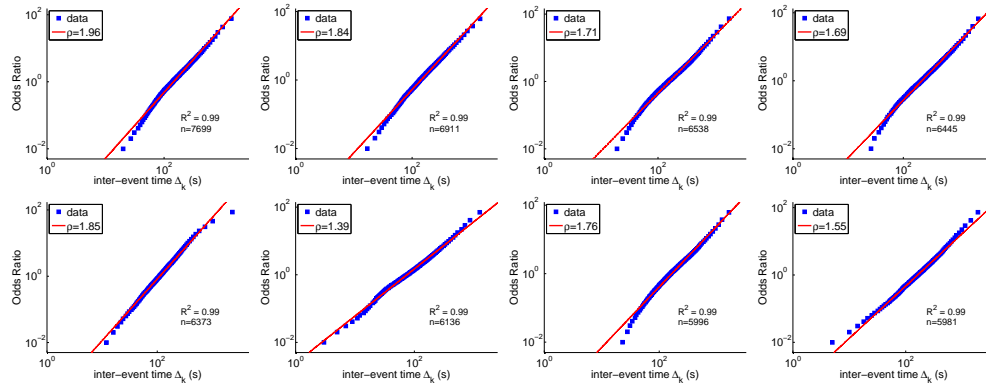


Fig. 25. Sample from the Phone dataset.

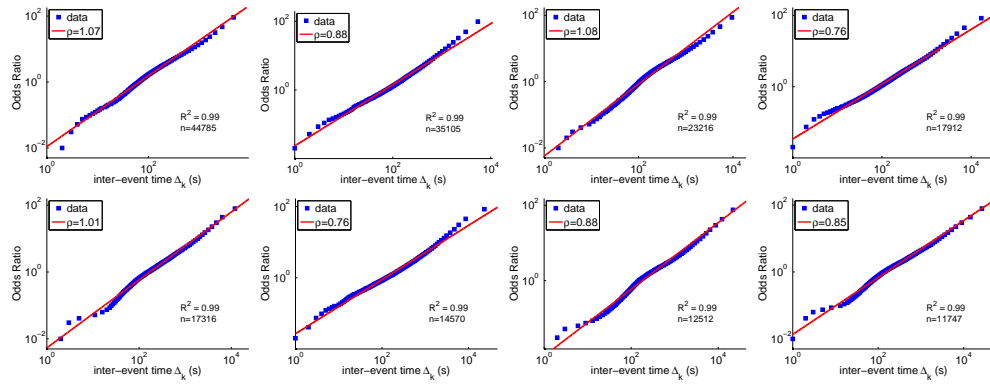


Fig. 26. Sample from the SMS dataset.

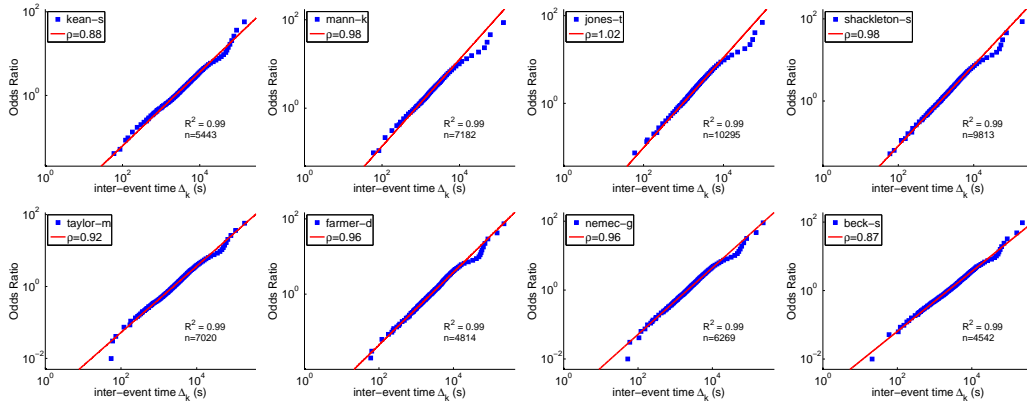


Fig. 27. Sample from the Enron dataset.

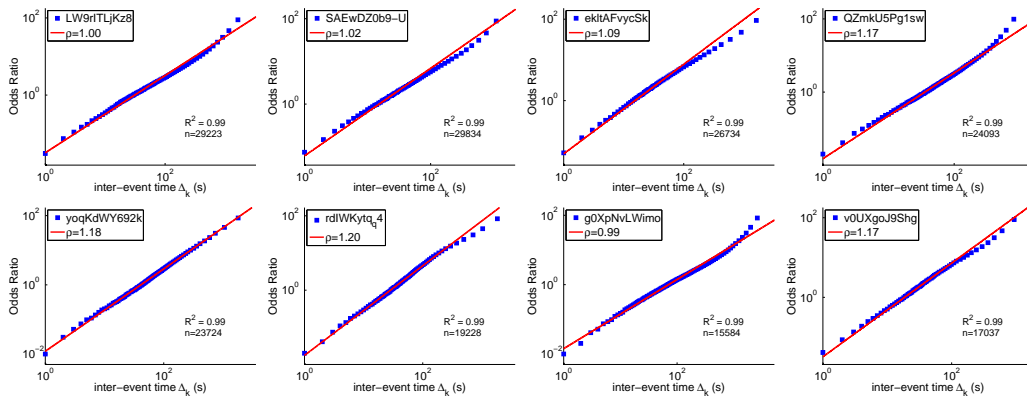


Fig. 28. Sample from the Youtube dataset.

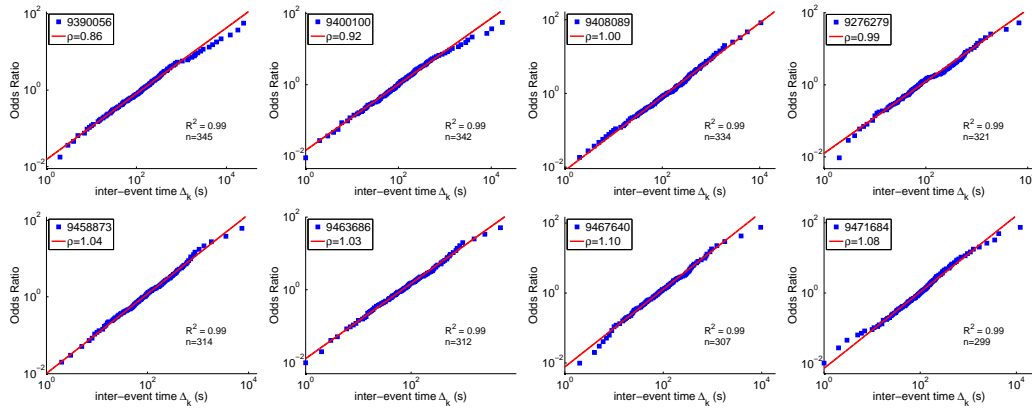


Fig. 29. Sample from the Digg dataset.

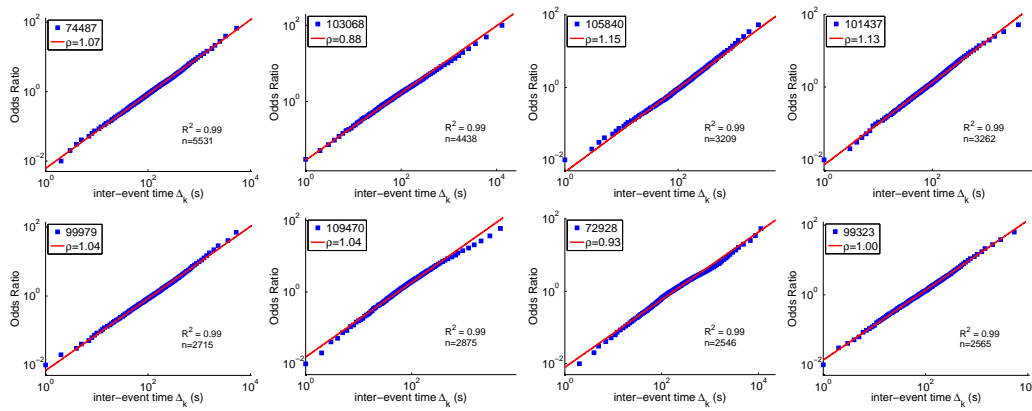


Fig. 30. Sample from the MetaFilter dataset.

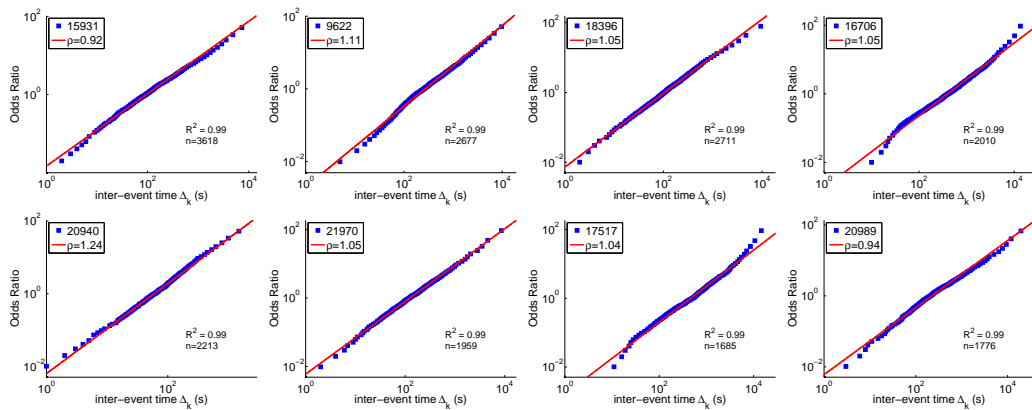


Fig. 31. Sample from the MetaTalk dataset.

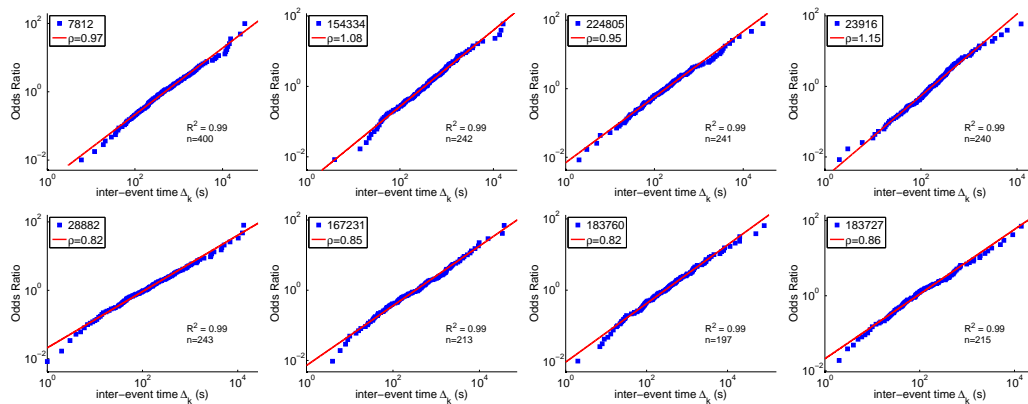


Fig. 32. Sample from the Ask MetaFilter dataset.

REFERENCES

- Leman Akoglu, Pedro O. S. Vaz de Melo, and Christos Faloutsos. 2012. Quantifying Reciprocity in Large Weighted Communication Networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2012, Kuala Lumpur*.
- A.-L. Barabási. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435 (16 May 2005), 207–211. DOI : <http://dx.doi.org/10.1038/nature03459>
- Steve Bennett. 1983. Log-Logistic Regression Models for Survival Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 32, 2 (1983), 165–171.
- G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. 1994. *Time Series Analysis, Forecasting, and Control* (third ed.). Prentice-Hall, Englewood Cliffs, New Jersey.
- Flavio Chierichetti, Ravi Kumar, Prabhakar Raghavan, and Tamas Sarlos. 2012. Are web users really Markovian?. In *Proceedings of the 21st international conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 609–618. DOI : <http://dx.doi.org/10.1145/2187836.2187919>
- Aaron Clauset, Cosma R. Shalizi, and M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Rev.* 51, 4 (2 Feb 2009), 661+. DOI : <http://dx.doi.org/10.1137/070710111>
- William S. Cleveland. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *J. Amer. Statist. Assoc.* 74, 368 (1979), 829–836. DOI : <http://dx.doi.org/10.2307/2286407>
- D.R. Cox and V. Isham. 1980. *Point Processes*. Taylor & Francis. <http://books.google.com.br/books?id=KWF2xY6s3PoC>
- D. R. Cox. 1955. Some Statistical Methods Connected with Series of Events. *Journal of the Royal Statistical Society. Series B (Methodological)* 17, 2 (1955), 129–164. DOI : <http://dx.doi.org/10.2307/2983950>
- Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. 2009. Social Synchrony: Predicting Mimicry of User Actions in Online Social Media. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*. IEEE Computer Society, Washington, DC, USA, 151–158. DOI : <http://dx.doi.org/10.1109/CSE.2009.439>
- Z. Dezső, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási. 2006. Dynamics of information access on the web. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* 73, 6 (2006), 066132. DOI : <http://dx.doi.org/10.1103/PhysRevE.73.066132>
- Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. 2004. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America* 101, 40 (5 October 2004), 14333–14337. DOI : <http://dx.doi.org/10.1073/pnas.0405728101>
- Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. 1999. On power-law relationships of the Internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*. ACM, New York, NY, USA, 251–262. DOI : <http://dx.doi.org/10.1145/316188.316229>
- Peter R. Fisk. 1961. The Graduation of Income Distributions. *Econometrica* 29, 2 (1961), 171–185.
- Swapna S. Gokhale and Kishor S. Trivedi. 1998. Log-Logistic Software Reliability Growth Model. In *HASE '98: The 3rd IEEE International Symposium on High-Assurance Systems Engineering*. IEEE Computer Society, Washington, DC, USA, 34–41.
- Frank A. Haight. 1967. *Handbook of the Poisson distribution [by] Frank A. Haight*. Wiley New York., xi, 168 p. pages.
- Uli Harder and Maya Paczuski. 2006. Correlated dynamics in human printing behaviour. *Physica A* 361, 1 (2006), 329–336.

- Cesar A. Hidalgo. 2006. Scaling in the Inter-Event Time of Random and Seasonal Systems. *PHYSICA A* 369 (2006), 877. <http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0512278>
- Richard A. Johnson and Dean W. Wichern. 2007. *Applied Multivariate Statistical Analysis (6th Edition)* (6 ed.). Pearson.
- T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido. 2004. A nonstationary Poisson view of Internet traffic, In *INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies. INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies* 3 (2004), 1558–1569 vol.3. DOI : <http://dx.doi.org/10.1109/INFCOM.2004.1354569>
- Márton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. 2012. Universal features of correlated bursty behaviour. *Scientific Reports* 2 (4 May 2012). DOI : <http://dx.doi.org/10.1038/srep00397>
- Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD (KDD '02)*. ACM, New York, NY, USA, 91–101. DOI : <http://dx.doi.org/10.1145/775047.775061>
- Bryan Klimt and Yiming Yang. 2004. Introducing the Enron Corpus. In *CEAS'04: The First Conference on Email and Anti-Spam* (2006-06-02). <http://dblp.uni-trier.de/db/conf/ceas/ceas2004.html#KlimtY04>
- A Kuczura. 1973. The interrupted Poisson process as an overflow process. *The Bell System Technical Journal* 52 (1973), 437–448.
- J. F. Lawless and Jerald F. Lawless. 1982. *Statistical Models and Methods for Lifetime Data (Wiley Series in Probability & Mathematical Statistics)*. John Wiley & Sons.
- M. O. Lorenz. 1905. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association* 9 (1905), 209–219.
- Talat Mahmood. 2000. Survival of Newly Founded Businesses: A Log-Logistic Model Approach. *Journal Small Business Economics* 14, 3 (2000), 223–237.
- R. Dean Malmgren, Jake M. Hofman, Luis A.N. Amaral, and Duncan J. Watts. 2009a. Characterizing individual communication patterns. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*. ACM, New York, NY, USA, 607–616. DOI : <http://dx.doi.org/10.1145/1557019.1557088>
- R. Dean Malmgren, Daniel B. Stouffer, Andriana S. L. O. Campanharo, and Luis A. Nunes Amaral. 2009b. On Universality in Human Correspondence Activity. *SCIENCE* 325 (2009), 1696. doi:10.1126/science.1174562
- R. Dean Malmgren, Daniel B. Stouffer, Adilson E. Motter, and Luís A. N. Amaral. 2008. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* 105, 47 (25 November 2008), 18153–18158. DOI : <http://dx.doi.org/10.1073/pnas.0800332105>
- C.D. Sinclair M.I. Ahmad and A. Werritty. 1988. Log-logistic flood frequency analysis. *Journal of Hydrology* 98 (1988), 205–224.
- Marcin Owczareczuk. 2011. Long memory in patterns of mobile phone usage. *Physica A: Statistical Mechanics and its Applications* (Oct. 2011). DOI : <http://dx.doi.org/10.1016/j.physa.2011.10.005>
- Pedro O. S. Vaz de Melo, Leman Akoglu, Christos Faloutsos, and Antonio Alfredo Ferreira Loureiro. 2010. Surprising Patterns for the Call Duration Distribution of Mobile Phone Users. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*. 354–369.
- Pedro Olmo Stancioli Vaz de Melo, C. Faloutsos, and A. A. Loureiro. 2011. Human Dynamics in Large Communication Networks. In *SIAM Conference on Data Mining (SDM)*. SIAM / Omnipress, 968–879.
- Alexei Vazquez, Joao Gama Oliveira, Zoltan Dezso, Kwang-Il Goh, Imre Kondor, and Albert-Lazlo Barabasi. 2006. Modeling bursts and heavy tails in human dynamics. *Phys Rev E Stat Nonlin Soft Matter Phys* 73 (2006), 036127.
- Hong Wei, Han Xiao-Pu, Zhou Tao, and Wang Bing-Hong. 2009. Heavy-Tailed Statistics in Short-Message Communication. *Chinese Physics Letters* 26, 2 (2009), 028902.
- H.O.A. Wold and Uppsala universitet. Statistiska institutionen. 1948. *On Stationary Point Processes and Markov Chains*. Swedish and Danish Actuarial Societies. <http://books.google.com.br/books?id=Mj5LNQAACAAJ>